Selective Unlearning in Neural Networks: Forgetting and Relearning Classes

October 26, 2024

Carmine Fabbri

Abstract

This project investigates machine unlearning by removing a specific class from a pre-trained neural network. We demonstrate this with an MNIST classifier aiming to forget class '6' and replace it with class '3'. The process involves identifying relevant weights, freezing irrelevant parameters, and fine-tuning the model using modified loss functions. We evaluate the model before and after unlearning through accuracy and confusion matrices, illustrating the effectiveness of our approach and potential challenges.

1. Introduction

As machine learning becomes prevalent in sensitive applications, such as data privacy and security, there is a pressing need for effective methods to remove specific information from pre-trained models. This process, known as **machine unlearning**, allows models to selectively forget information, including private or outdated data, without compromising overall performance.

In this project, a custom-trained model is fine-tuned by freezing non-essential parameters and using a modified loss function. The evaluation incorporates both quantitative (accuracy) and qualitative (confusion matrix) metrics to demonstrate the impact of selective unlearning on the model's performance.

2. Related Work

Selective unlearning has been explored in various contexts, particularly in privacy-preserving machine learning. One notable approach is *Elastic Weight Consolidation (EWC)*, introduced by Kirkpatrick et al. (2017), which protects critical weights during retraining to prevent catastrophic forgetting in lifelong learning scenarios.

Email: Carmine Fabbri <fabbri.2133421@studenti.uniroma1.it>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

Other research, such as that by *Golatkar et al. (2020)*, focuses on explicit unlearning techniques designed to mitigate the influence of individual data points in neural networks. These methods are particularly valuable when data must be removed due to regulatory requirements or privacy issues.

In contrast to EWC's goal of retaining knowledge while learning new tasks, our work emphasizes **selective forgetting**, specifically targeting the removal of class information.

3. Methodology

3.1. Model architecture

The project employs a custom Convolutional Neural Network (CNN) for MNIST digit classification, featuring:

- **Convolutional Layers**:
  - **conv1**: Processes single-channel (grayscale) input with 32 filters of size 3×3, stride 1, and padding 1.
  - **conv2**: Applies 64 filters of size 3×3 with the same stride and padding.
  - Each convolutional layer is followed by a **ReLU** activation and a **2 x 2 max-pooling** layer, reducing the spatial dimensions to 7×7.
- **Fully Connected Layers**:
  - **fc1**: Flattens the output to 3,136 units and maps to 128 units.
  - **fc2**: Reduces 128 units to 64 units.
  - **fc3**: Outputs 10 units, corresponding to the MNIST digit classes.
- **Output Layer**: Utilizes a **log-softmax** function to generate a probability distribution over the 10 classes.

This architecture effectively combines feature extraction and decision-making, making it suitable for image classification tasks.

3.2. Selective Unlearning Process

The unlearning process aims to remove the model's ability to correctly classify instances of digit 6 while retaining performance on other digits, particularly class 3. This is done using the following steps:

1. **Identifying Relevant Weights**: During backpropagation, the weights associated with class 6 are identified by inspecting the gradients and activations.
2. **Freezing Non-Relevant Parameters**: To prevent the model from forgetting other classes, all weights not heavily involved in class 6 predictions are frozen. This allows the model to focus only on modifying the relevant weights during unlearning.
3. **Modified Loss Function**: A custom loss function is crafted to selectively penalize and guide the model's predictions. It imposes a penalty on predictions for class 6, by discouraging the model from associating data with this class. In contrast, the function encourages accurate predictions for class 3 by applying standard cross-entropy loss to other classes.

3.3. Evaluation Metrics

The model's performance is assessed using two main metrics: **Accuracy**, which measures the overall performance before and after the unlearning process, and the **Confusion Matrix**, which evaluates classification accuracy for each digit, with a focus on changes in the classifications of classes 6 and 3.

Additionally, custom-created plotting functions were used to visualize model accuracy and loss over epochs, highlighting performance throughout training and unlearning and illustrating convergence during fine-tuning.

4. Experimental Results

4.1. Quantitative Results

Before the unlearning process, the model achieved an accuracy of 98.57% on the MNIST test set. After unlearning class 6 and fine-tuning for class 3, the accuracy decreased to 89.36%, which was anticipated due to the intentional removal of class 6. However, performance on class 3 improved slightly, demonstrating successful relearning.

Metrics	Before Unlearning	After Unlearning
Overall Accuracy	98.57%	89.36%
Class 3 Accuracy	97.43%	99.01%
Class 6 Accuracy	99.27%	0.00%

Table 1. Accuracy Metrics Before and After Unlearning

4.2. Qualitative Results

- **Confusion Matrix Before Unlearning:** Showed high accuracy across all classes, including class 6, with minimal misclassifications.

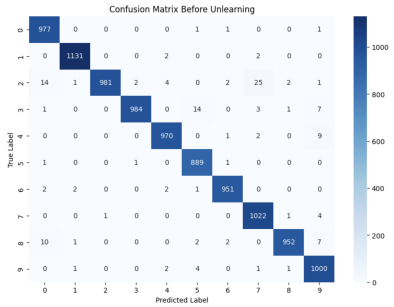

Figure 1. Confusion Matrix Before Unlearning

- **Confusion Matrix After Unlearning:** Indicated that instances of class 6 were misclassified as other digits, particularly class 3. This result suggests that the model effectively "forgot" class 6 while maintaining or improving predictions for other classes.

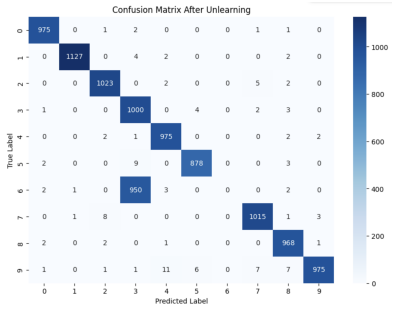

Figure 2. Confusion Matrix After Unlearning

5. Discussion and Conclusion

The results demonstrate that selective unlearning can be applied effectively in neural networks by penalizing specific class predictions and freezing non-relevant weights. The primary challenge is preventing *catastrophic forgetting*, which can cause the model to lose performance on classes unrelated to the unlearning target. However, our approach showed that selective unlearning can be done with minimal loss in performance on other classes.

In conclusion, this project demonstrates the potential of selective unlearning in machine learning models, with promising applications in privacy-preserving machine learning and incremental learning. However, care must be taken to mitigate performance degradation in the overall model while unlearning specific classes.

Bibliography. (Kirkpatrick et al., 2017).
(Golatkar et al., 2020)

References

Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 2017.